New techniques for best-match retrieval

Full text Pdf (1.41 MB)
Source ACM Transactions on Information Systems (TOIS) archive
Volume 8 , Issue 2 (April 1990) table of contents
Pages: 140 - 158
Year of Publication: 1990
ISSN:1046-8188

Authors Dennis Shasha New York Univ., New York
Tsong-Li Wang New York Univ., New York

Publisher ACM New York, NY, USA

ABSTRACT

A scheme to answer best-match queries from a file containing a collection of objects is described. A best-match query is to find the objects in the file that are closest (according to some (dis)similarity measure) to a given target. Previous work [5, 331] suggests that one can reduce the number of comparisons required to achieve the desired results using the triangle inequality, starting with a data structure for the file that reflects some precomputed intrafile distances. We generalize the technique to allow the optimum use of any given set of precomputed intrafile distances. Some empirical results are presented which illustrate the effectiveness of our scheme, and its performance relative to previous algorithms.

REFERENCES

Note: OCR errors may be found in this Reference List extracted from the full text article. ACM has opted to expose the complete List rather than only correct and linked references.

1 Rakesh Agrawal , H. V. Jagadish, Efficient Search in Very Large Databases, Proceedings of the 14th International Conference on Very Large Data Bases, p.407-418, August 29-September 01, 1988

2 Alfred V. Aho , John E. Hopcroft , Jeffrey Ullman , J. D. Ullman , J. E. Hopcroft, Data Structures and Algorithms, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, 1983

3 D. P. Anderson, Techniques for Reducing Pen Plotting Time, ACM Transactions on Graphics (TOG), v.2 n.3, p.197-212, July 1983 [doi> 10.1145/357323.357327]

4 Jon Louis Bentley , Bruce W. Weide , Andrew C. Yao, Optimal Expected-Time Algorithms for Closest Point Problems, ACM Transactions on Mathematical Software (TOMS), v.6 n.4, p.563-580, Dec. 1980 [doi> 10.1145/355921.355927]

5 W. A. Burkhard , R. M. Keller, Some approaches to best-match file searching, Communications of the ACM, v.16 n.4, p.230-236, April 1973 [doi> 10.1145/362003.362025]

6 CL~,US, V., EHRIG, M., AND ROZENBERG, G. Graph-Grammars and Their Application to Computer Science and Biology. Springer, New York, 1979.

7 Du, H. C., AND LEE, R. C.W. Symbolic Gray code as a multikey hashing function. IEEE Trans. Pattern Anal. Mach. InteU. 2, 1 (Jan. 1980), 83-90.

8 DUDA, R. O., AND HART, P.E. Pattern Classification and Scene Analysis. Wiley, New York, 1973.

9 Caroline M. Eastman , Stephen F. Weiss, A tree algorithm for nearest neighbor searching in document retrieval systems, Proceedings of the 1st annual international ACM SIGIR conference on Information storage and retrieval, p.131-149, May 10-12, 1978

10 EASTMAN, C. M., AND WEISS, S.F. Tree structures for high dimensionality nearest neighbor searching. Inf. Syst. 7, 2 (1982), 115-122.

11 EASTMAN, C. M., AND ZEMANKOVA, M. Partially specified nearest neighbor searches using k - d trees. Inf. Process. Lett. 15, 2 (1982), 53-56.

12 FEUSTEL, C. D., AND SHAPIRO, L.G. The nearest neighbor problem in an abstract metric space. Pattern Recognition Lett. 1, 2 (1982), 125-128.

13 Robert W. Floyd, Algorithm 97: Shortest path, Communications of the ACM, v.5 n.6, p.345, June 1962 [doi> 10.1145/367766.368168]

14 Jerome H. Freidman , Jon Louis Bentley , Raphael Ari Finkel, An Algorithm for Finding Best Matches in Logarithmic Expected Time, ACM Transactions on Mathematical Software (TOMS), v.3 n.3, p.209-226, Sept. 1977 [doi> 10.1145/355744.355745]

15 FUKUNAGA, K., AND NARENDRA, P.M. A branch and bound algorithm for computing k-nearest neighbors. IEEE Trans. Comput. 24, 7 (July 1975), 750-753.

16 HOLLAAR, L.A. The Utah text retrieval project. Inf. Technol. Res. Dev. 2 (1983), 155-167.

17 Tetsuro Ito , Makoto Kizawa, Hierarchical file organization and its application to similar-string matching, ACM Transactions on Database Systems (TODS), v.8 n.3, p.410-433, Sept. 1983 [doi> 10.1145/319989.319994]

18 LANDER, E., MESIROV, J. P., AND WASHINGTON, T. Protein sequence comparison on a data parallel computer. In Proceedings of the IEEE 1988 International Conference on Parallel Processing (Aug. 1988). IEEE, New York, 1988, 257-263.

19 LIPMAN, D. J., AND PEARSON, W.R. Rapid and sensitive protein similarity searches. Science 227 (1985), 1435-1441.

20 Dario Lucarella, A document retrieval system based on nearest neighbour searching, Journal of Information Science, v.14 n.1, p.25-33, January 1988 [doi> 10.1177/016555158801400104]

21 K C Mohan , P Willett, Nearest neighbour searching in serial files using text signature,

Journal of Information Science, v.11 n.1, p.31-39, 1985
[doi>10.1177/016555158501100105]

22 MURTAGH, F. A very fast exact nearest neighbor algorithm for use in information retrieval. Inf. Technol. Res. Dev. I (1982), 275-283.

23 MURTAGH, F. Expected-time complexity results for hierarchic clustering algorithms which use cluster centers. Inf. Process. Lett. 16, 5 (June 1983), 237-241.

24 MURTAGH, F. A survey of recent advances in hierarchical clustering algorithms. IEEE Computer 26, 4 (1983), 354-359.

25 MURTAGH, F. Multidimensional clustering algorithms. In Lectures in Computational Statistics, J. M. Chambers, J. Gordesch, A. Klas, L. Lebart, and P. P. Sint, Eds., Physica-Verlag, Vienna, 1985.

26 PA{ GE, R. C., AND Kr~ USKAL, C.P. Parallel algorithms for shortest path problems. In Proceedings of the IEEE 1985 International Conference on Parallel Processing (1985). IEEE, New York, 1985, 14-19.

27 PERRY, S. A., AND WILLETT, P. A review of the use of inverted files for best match searching in information retrieval systems. J. Inf. Sci. 6 (1983), 59-66.

28 POOUE, C. A., AND WILLETT, P. An evaluation of document retrieval from serial files using the ICL Distributed Array Processor. Online Rev. 8 (1984), 569-584.

29 ROHLF, F.J. A probabilistic minimum spanning tree algorithm, inf. Process. Lett. 7 {1978), 44-48.

30 Gerard Salton , Michael J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, Inc., New York, NY, 1986

31 SHAMOS, M. I., AND HOE~ , D. Closest-point problems. In Proceedings of the 16th IEEE Symposium on Foundations of Computer Science (Oct. 1975). IEEE, New York, 1975, 151-162.

32 SHAPIRO, B. A., AND ZHANG, K. Comparing multiple RNA secondary structures using tree comparisons. Manuscript, Division of Cancer Biology and Diagnosis, NIH, Frederick, Md., 1989.

33 Marvin Shapiro, The choice of reference points in best-match file searching, Communications of the ACM, v.20 n.5, p.339-343, May 1977
[doi>10.1145/359581.359599]

34 SHASHA, D., AND WANG, T.-L. Optimal best-match retrieval. Tech. Rep. TR 480, Courant Institute of Mathematical Sciences, New York Univ., New York, Dec. 1989.

35 A. F. Smeaton , C. J. van Rijsbergen, The nearest neighbour problem in information retrieval: an algorithm using upperbounds, ACM SIGIR Forum, v.16 n.1, p.83-87, Summer 1981

36 Mark Stewart , Peter Willett, Nearest neighbour searching in binary search trees: Simulation of a multiprocessor system, Journal of Documentation, v.43 n.2, p.93-111, June 1987 [doi>10.1108/eb026803]

37 TESKEY, F.N. Novel computer architectures for data storage and retrieval. Re/). 5845, British Library Research and Development Dept., London, 1986.

38 C. J. van Rijsbergen, The best-match problem in document retrieval, Communications of the ACM, v.17 n.11, p.648-649, Nov. 1974 [doi>10.1145/361179.361205]

39 C. J. Van Rijsbergen, Information Retrieval, Butterworth-Heinemann, Newton, MA, 1979

40 Tsong-Li Wang , Dennis Shasha, Query processing for distance metrics, Proceedings of the sixteenth international conference on Very large databases, p.602-613, September 1990, Brisbane, Australia

41 Stephen Warshall, A Theorem on Boolean Matrices, Journal of the ACM (JACM), v.9 n.1, p.11-12, Jan. 1962 [doi>10.1145/321105.321107]

42 C. T. Yu , W. S. Luk , M. K. Siu, On the estimation of the number of desired records with respect to a given query, ACM Transactions on Database Systems (TODS), v.3 n.1, p.41-56, March 1978 [doi>10.1145/320241.320245]

CITED BY *18*

Sergey Brin, Near Neighbor Search in Large Metric Spaces, Proceedings of the 21th International Conference on Very Large Data Bases, p.574-584, September 11-15, 1995

Pierre Zakarauskas , John M. Ozard, Complexity Analysis for Partitioning Nearest Neighbor Searching Algorithms, IEEE Transactions on Pattern Analysis and Machine Intelligence, v.18 n.6, p.663-668, June 1996

Tolga Bozkaya , Meral Ozsoyoglu, Distance-based indexing for high-dimensional metric spaces, ACM SIGMOD Record, v.26 n.2, p.357-368, June 1997

Ulrich Pfeifer , Norbert Fuhr, Efficient processing of vague queries using a data stream approach, Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval, p.189-197, July 09-13, 1995, Seattle, Washington, United States

Shian-Hua Lin , Chi-Sheng Shih , Meng Chang Chen , Jan-Ming Ho , Ming-Tat Ko , Yueh-Ming Huang, Extracting classification knowledge of Internet documents with mining term associations: a semantic approach, Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, p.241-249, August 24-28, 1998, Melbourne, Australia

Shian-Hua Lin , Jan-Ming Ho, Discovering informative content blocks from Web documents, Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, July 23-26, 2002, Edmonton, Alberta, Canada

Alfredo Ferro , Giovanni Gallo , Rosalba Giugno , Alfredo Pulvirenti, Best-Match Retrieval for Structured Images, IEEE Transactions on Pattern Analysis and Machine Intelligence, v.23 n.7, p.707-718, July 2001

Edwin M. Knorr , Raymond T. Ng, Finding Aggregate Proximity Relationships and Commonalities in Spatial Data Mining, IEEE Transactions on Knowledge and Data Engineering, v.8 n.6, p.884-897, December 1996

Tolga Bozkaya , Meral Ozsoyoglu, Indexing large metric spaces for similarity search queries, ACM Transactions on Database Systems (TODS), v.24 n.3, p.361-404, Sept. 1999

C. Traina, Jr. , A. Traina , C. Faloutsos , B. Seeger, Fast Indexing and Visualization of Metric Data Sets using Slim-Trees, IEEE Transactions on Knowledge and Data Engineering, v.14 n.2, p.244-260, March 2002

Jason Tsong-Li Wang , Gung-Wei Chirn , Thomas G. Marr , Bruce Shapiro , Dennis Shasha , Kaizhong Zhang, Combinatorial pattern discovery for scientific data: some preliminary results, ACM SIGMOD Record, v.23 n.2, p.115-125, June 1994

J. T. -L. Wang , K. Zhang , K. Jeong , D. Shasha, A System for Approximate Tree Matching,

IEEE Transactions on Knowledge and Data Engineering, v.6 n.4, p.559-571, August 1994

S.-H. Lin , M. C. Chen , J.-M. Ho , Y.-M. Huang, ACIRD: Intelligent Internet Document Organization and Retrieval, IEEE Transactions on Knowledge and Data Engineering, v.14 n.3, p.599-614, May 2002

Flip Korn , Bernd-Uwe Pagel , Christos Faloutsos, On the 'Dimensionality Curse' and the 'Self-Similarity Blessing', IEEE Transactions on Knowledge and Data Engineering, v.13 n.1, p.96-111, January 2001

Domenico Cantone , Alfredo Ferro , Alfredo Pulvirenti , Diego Reforgiato Recupero , Dennis Shasha, Antipole Tree Indexing to Support Range Search and K-Nearest Neighbor Search in Metric Spaces, IEEE Transactions on Knowledge and Data Engineering, v.17 n.4, p.535-550, April 2005

Christos Faloutsos , King-Ip Lin, FastMap: a fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets, ACM SIGMOD Record, v.24 n.2, p.163-174, May 1995

Edgar Chávez , Gonzalo Navarro , Ricardo Baeza-Yates , José Luis Marroquín, Searching in metric spaces, ACM Computing Surveys (CSUR), v.33 n.3, p.273-321, September 2001

Friedrich Gebhardt, Survey on structure-based case retrieval, The Knowledge Engineering Review, v.12 n.1, p.41-58, January 1997

↑ INDEX TERMS

Primary Classification:
- E. Data
  - E.5 FILES
    - Subjects: Sorting/searching

Additional Classification:
- F. Theory of Computation
  - F.2 ANALYSIS OF ALGORITHMS AND PROBLEM COMPLEXITY
    - F.2.2 Nonnumerical Algorithms and Problems
      - Subjects: Sorting and searching
- H. Information Systems
  - H.2 DATABASE MANAGEMENT
    - H.2.2 Physical Design
      - Subjects: Access methods
    - H.2.4 Systems
      - Subjects: Query processing
  - H.3 INFORMATION STORAGE AND RETRIEVAL
    - H.3.3 Information Search and Retrieval
      - Subjects: Search process

General Terms:
Algorithms, Design, Experimentation, Performance

REVIEW

"Caroline Merriam Eastman : Reviewer"

The best-match problem involves finding the objects in a file that are closest to some specified object. Two major issues are the choice of an appropriate measure of closeness and efficient implementation. This paper addresses the *more...*

Collaborative Colleagues:

Dennis Shasha: colleagues

Tsong-Li Wang: colleagues

Peer to Peer - Readers of this Article have also read:

- M⁴: a metamodel for data preprocessing Proceedings of the 4th ACM international workshop on Data warehousing and OLAP
  Anca Vaduva , Jörg-Uwe Kietz , Regina Zücker

- Data structures for quadtree approximation and compression Communications of the ACM 28, 9
  Hanan Samet

- A hierarchical single-key-lock access control using the Chinese remainder theorem Proceedings of the 1992 ACM/ SIGAPP Symposium on Applied computing
  Kim S. Lee , Huizhu Lu , D. D. Fisher

- The GemStone object database management system Communications of the ACM 34, 10
  Paul Butterworth , Allen Otis , Jacob Stein

- Putting innovation to work: adoption strategies for multimedia communication systems Communications of the ACM 34, 12
  Ellen Francik , Susan Ehrlich Rudman , Donna Cooper , Stephen Levine